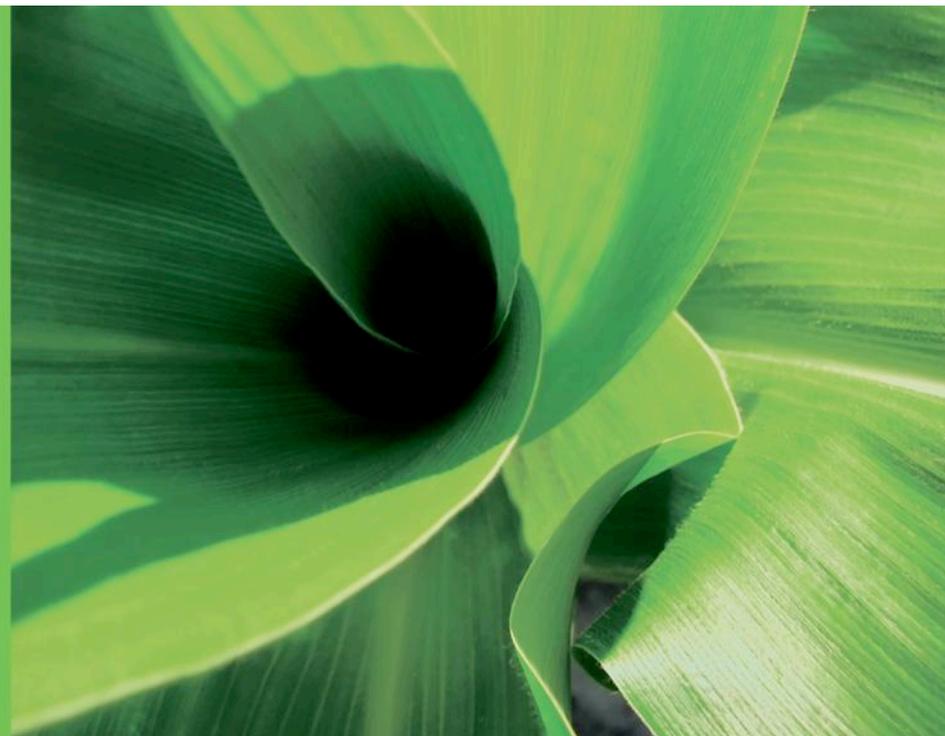


# High Throughput Plasmid Sequencing with Illumina and CLC bio

Ajay Athavale  
Monsanto Company  
05 June 2012



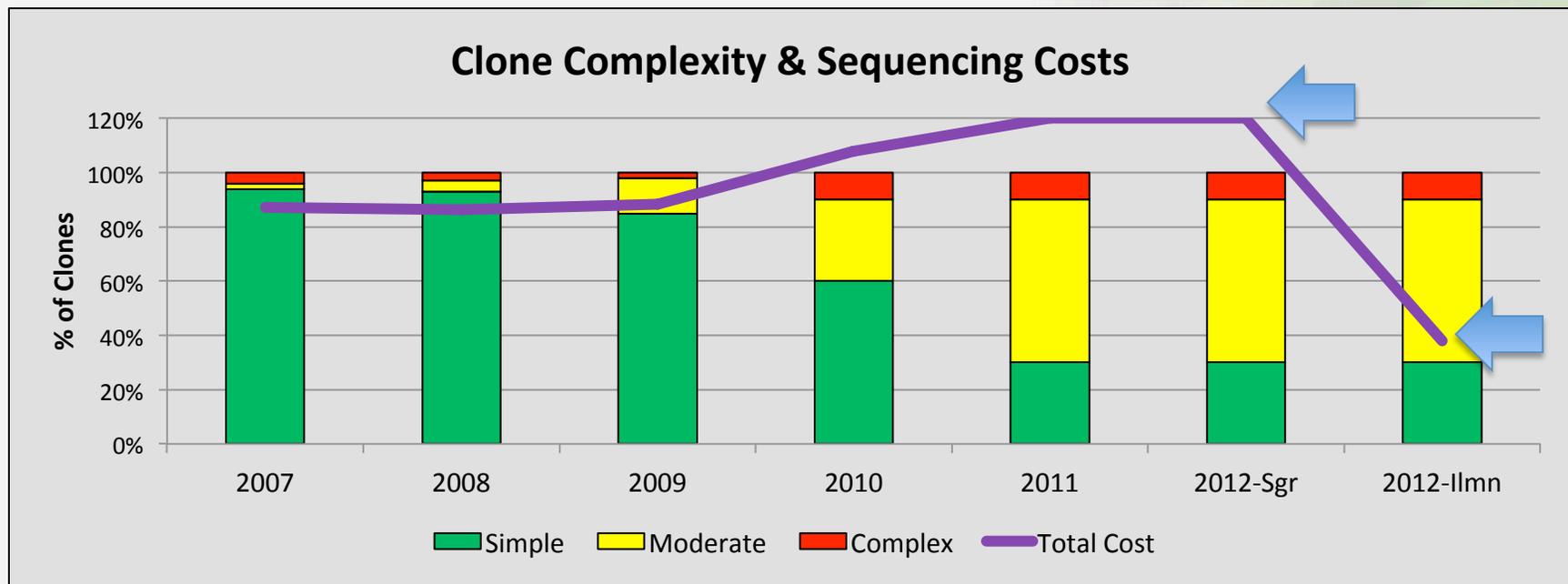
6/4/2012

MONSANTO  Monsanto Company Confidential

# Monsanto's Transgenic Plant Pipeline: Sequencing 1,000's of plasmids per year



# Increasing Clone Complexity Required New Solutions for Plasmid Sequencing & Analysis



- NextGen sequencing offers significant savings & reduced processing time
- Illumina was demonstrated to be optimal NGS platform for plasmid sequencing
- Scalable and accurate finishing tools were needed for Illumina sequence reads

# Requirements for HTP Plasmid Finishing Platform

## Performance, Accuracy & Scalability

- Support structurally challenging plasmids
- Support for multiple sequencing platforms
- Hybrid assemblies (across platforms)
- User configurable & flexible parameters
  - Read trimming
  - Mutation detection (SNPs)
  - Insertion & Deletion detection (DIPs)

## Usability & User Interface

- Easily edit/manipulate assemblies
- Execution from both UNIX and GUI environments
- User friendly visualization tools

## Integration with in house corporate research data systems

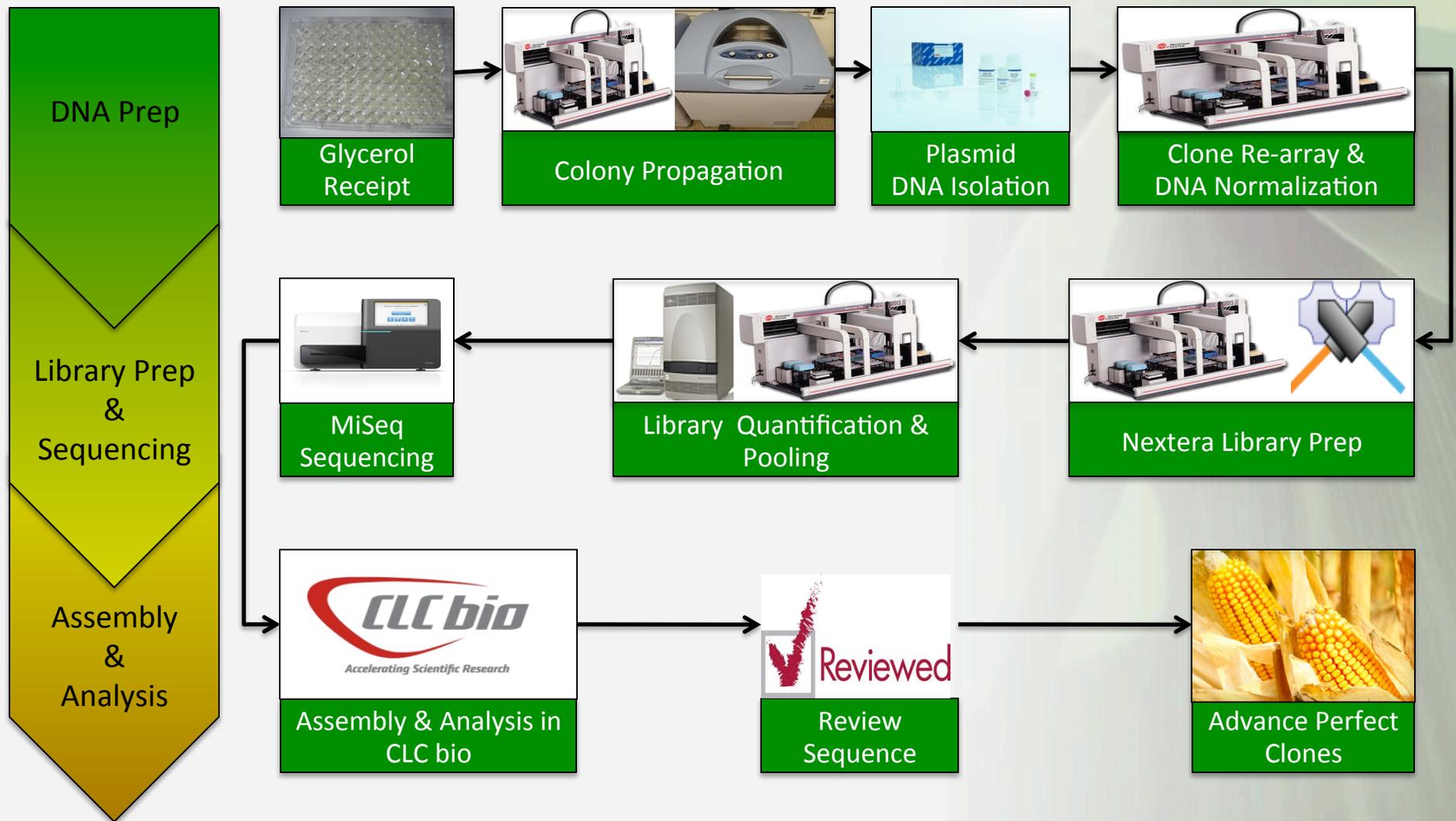
## Customization of finishing workflows

 **CLC bio was chosen as next gen finishing tool, given these criteria**

## Demonstrated identical finishing results for structurally complex plasmids using Illumina and CLC Bio

Plasmid	Known Complex Sequences	Observed Seq Sanger + CONSED	Observed Seq Illumina + CLC bio
Plasmid A	None	☑	☑
Plasmid B	Mixed Population	☑	☑
Plasmid C	Homopolymer tracts	☑	☑
Plasmid D	Homopolymer tracts Tandem Repeats	☑	☑
Plasmid E	Homopolymer tracts Inverted Repeats	☑	☑
Plasmid F	Homopolymer tracts Inverted Repeats Tandem Repeats	☑	☑
Plasmid G	Large Insertion	☑	☑
Plasmid H	Large Deletion	☑	☑

# Overview of the Current Sequencing & Finishing Process with Illumina and CLC bio



# Automated Assembly & Analysis Workflow in CLC bio streamlines workflow

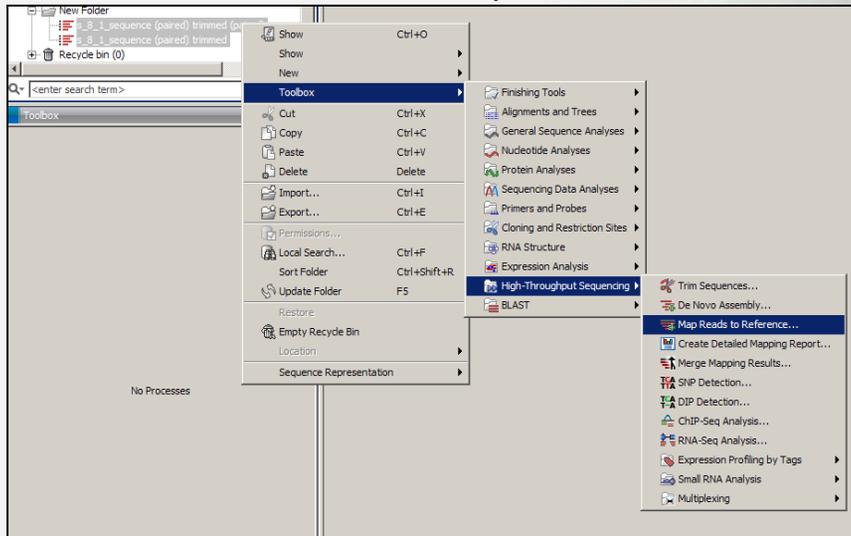
## Assembly & Analysis Workflow



- The CLC bio Genomics Server and Command line tools enable efficient parallel processing of large batches of constructs

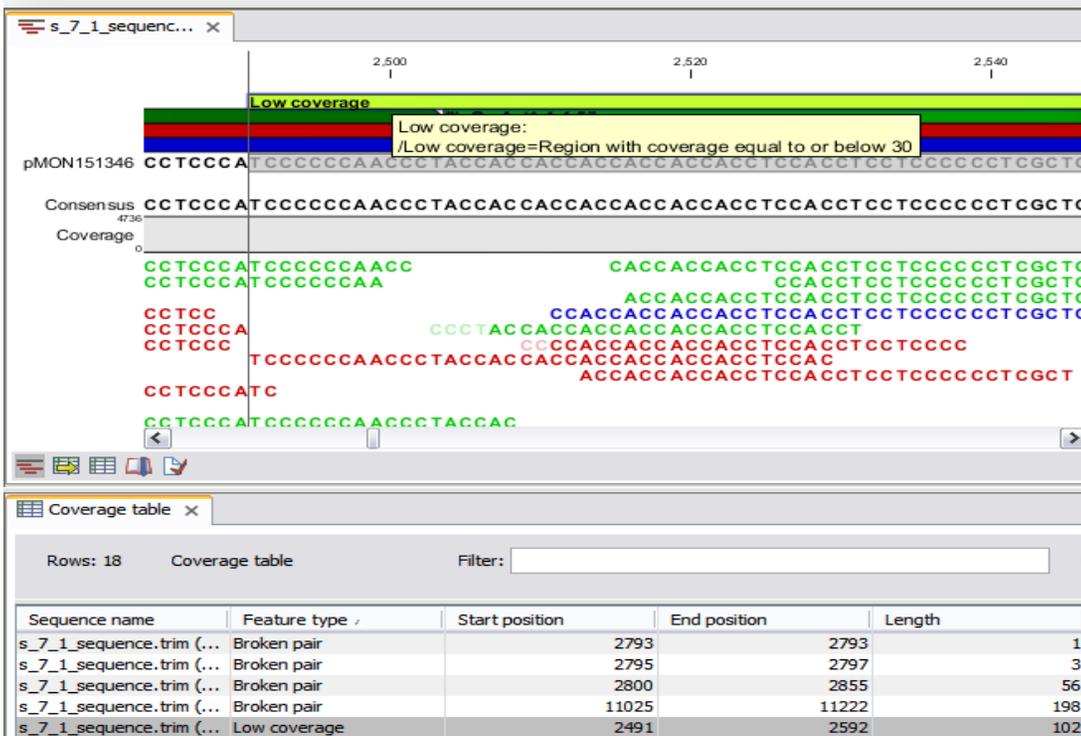
```
unix:> ./clserver -S server -U user -W password -A  
read mapping -i input reads --references reference  
-d save destination
```

- For the non-UNIX user, command line tools can be easily run in the GUI (CLC Genomics workbench)

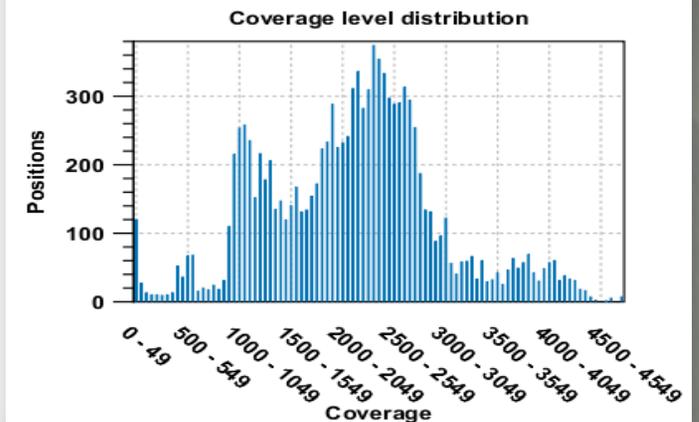


# Identification of low coverage regions with the “Contig Analysis” tool

Easy identification of low coverage areas



## 2.3 Coverage level distribution



## 2.2 Coverage statistics

Total reference length	11,222
Minimum coverage	6
Maximum coverage	4,736
Average coverage	2,118.12
Standard deviation	855.63

# The “Primer Creator” tool helps to streamline efforts to close gaps

The image displays the Primer Creator tool interface. At the top, a sequence alignment for pMON151346 is shown with a 'Low coverage' region highlighted in green. A tooltip explains: 'Low coverage: /Low coverage=Region with coverage equal to or below 30'. Below this, the 'Primer Creator' window is open, showing two steps: '1. Choose where to run' and '2. Select one or more nucleotide sequences'. The 'Set parameters' section includes 'Mispriming parameters' with a checked option 'Use mispriming as exclusion criteria'. The main window shows a sequence alignment with a 'Low coverage' region highlighted in green. A forward primer (Fwd) is selected at position 2457-2475, and a reverse primer (Rev) is selected at position 2605-2622. Below the alignment, a primer table is displayed:

Sequence ...	Primer seq...	Start position	End position	Length	Score	Direction	Melting temperature	GC content	Self annealing	Self end annealing
pMON151346	CGCCACTA...	2457	2475	19	41	Forward	53.40	47.37	18	3
pMON151346	AAGAAAGA...	2605	2622	18	43	Reverse	57.36	55.56	6	6





# Denovo assembly used to resolve Indels

The screenshot displays the 'De Novo Assembly' software interface. At the top, there are two window titles: 'g. De Novo Assembly' and 'g. De Novo Assembly'. The main window shows a list of steps: 1. Choose where to run, 2. Select sequencing reads, and 3. [unclear]. Below the steps, there are two sub-panels: 'Select de novo options' and 'Graph parameters'. A table with 658 rows is displayed, with a filter box above it. The table has the following columns: Name, Consensus length, Total read count, Single reads, Reads in pairs, and Average coverage. The first row is highlighted in green. Below the table, there are several checkboxes: 'Perform scaffolding' (checked), 'Update contigs' (checked), and 'Create list of un-mapped reads' (unchecked). At the bottom, there are navigation buttons: 'Previous', 'Next', 'Finish', and 'Cancel'. A yellow warning icon is visible in the bottom left corner.

Name	Consensus length	Total read count	Single reads	Reads in pairs	Average coverage
s_1_1_sequence (...)	10819	62002	11160	50842	697.31
s_1_1_sequence (...)	595	9	3	6	2.07
s_1_1_sequence (...)	596	9	1	8	2.17
s_1_1_sequence (...)	259	6	0	6	3.08
s_1_1_sequence (...)	339	12	8	4	2.72
s_1_1_sequence (...)	276	4	2	2	2.05
s_1_1_sequence (...)	292	5	1	4	2.19

typically

# CLC Bio Meets Requirements for Monsanto HTP Plasmid Finishing

## ✓ Performance, Accuracy & Scalability

- ✓ Support structurally challenging plasmids
- ✓ Support for multiple sequencing platforms
- ✓ Hybrid assemblies (across platforms)
- ✓ User configurable & flexible parameters
  - Read trimming
  - Mutation detection (SNPs)
  - Insertion & Deletion detection (DIPs)

## ✓ Usability & User Interface

- ✓ Easily edit/manipulate assemblies
- ✓ Execution from both UNIX and GUI environments
- ✓ User friendly visualization tools

## ✓ Integration with in house corporate research data systems

## ✓ Customization of finishing workflows

## **Additional CLC bio capabilities are also used across Monsanto**

- In silico cloning & plasmid design
- Sanger sequence review/assembly
- Plasmid sequencing & finishing via Sanger & Illumina
- Genome assembly for microbes and plants
- DNA and Protein alignment for research teams
- ... and much more!

# Acknowledgements

- CLC Bio
  - Jannick Bendtsen
  - Henrik Sandmann
  - Joe Salvatore
- Monsanto
  - Todd Michael
  - Dan Ader
  - Amber Ford
  - Susan Johnson
  - Cynthia LaBanca
  - Kim Lawry
  - Jing Lu
  - Karen Martin
  - Tim Mitsky
  - Stacie Norton